UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/577,174	04/26/2006	Masashi Inoue	JP920030230US1	1410

30449 7590 03/18/2008

SCHMEISER, OLSEN & WATTS
22 CENTURY HILL DRIVE
SUITE 302
LATHAM, NY 12110

EXAMINER	
GUPTA, MUKTESH G	
ART UNIT	PAPER NUMBER
2144	
MAIL DATE	DELIVERY MODE
03/18/2008	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	Application No.	Applicant(s)	
	10/577,174	INOUE, MASASHI	
	Examiner	Art Unit	
	Muktesh G. Gupta	2144	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE *3* MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *26 April 2006*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-38* is/are pending in the application.

4a) Of the above claim(s) *1-14* is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *15-38* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *26 April 2006* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☒ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☒ All b)☐ Some * c)☐ None of:

1.☒ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date *04/26/2006*.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

1. **Claims 1-14,** have been cancelled in this application.

2. **Claims 15-38,** have been examined on merits and are pending in this application.

Information Disclosure Statement

3. The information disclosure statement (IDS) submitted on 04/26/2006 is being considered by the examiner.

Priority

4. Acknowledgment is made of applicant's claim for foreign priority under 35 U.S.C. 119(a)-(d). The certified copy has been filed in parent Application No. 10/577174 filed on 04/26/2006.

Claim Rejections - 35 USC § 102

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (e) the invention was described in (1) an application for patent, published under section 122(b), by another filed in the United States before the invention by the applicant for patent or (2) a patent granted on an application for patent by another filed in the United States before the invention by the applicant for patent, except that an international application filed under the treaty defined in section 351(a) shall have the effects for purposes of this subsection of an application filed in the United States only if the international application designated the United States and was published under Article 21(2) of such treaty in the English language.

5. Claims 15-38, rejected under 35 U.S.C. 102(e) as being anticipated by U.S. Patent No. 7146353 to Garg, Pankaj K. et al., (hereinafter "Garg").

As to Claims 15, 24 and 32, *Garg teaches method, Load Control Server and Computer readable medium, for controlling bottlenecks in an information system that includes N application servers and a database server, wherein* (as stated in col. 1, lines 53-54, col. 4, lines 44-45, col. 5, lines 15-21, various ***methods*** and ***apparatus*** for efficiently allocating ***resources*** which are automatically reconfigured for ***optimizing requirements*** and adjusting ***load balancing*** policies to ***plurality*** of ***applications***. Where the system is modeled as an open queuing network, with three tiers, ***web servers*** to an ***application servers*** to a ***database servers*** arranged in *series, and parallel*, identical servers within each tier)*:*

N is at least 2, wherein each application server is adapted to execute at least one application program for processing a transaction received by each application server from a terminal, wherein (as stated in col. 3, lines 5-14, ***Clients*** access the ***applications*** via a network such as the Internet 120, and ***applications requirements*** are being satisfied by allocating ***resources efficiently***, various ***transaction data*** is collected (step 202). This transaction or instrumentation data may include data that identify transactions, classify transactions, identify ***requesters***, and quantify responsiveness of various ***components involved*** in ***processing*** the ***transaction***.

Data collected at various components involved in ***processing the transactions*** may be ***correlated*** by ***transactions***);

the database server is adapted to access a database based on a request received from any application server of the N application servers, and wherein the method comprises (as stated in col. 5, lines 18-21, multiple-pass ***processing*** of returned ***requests*** is ***aggregated*** into a one-pass ***simplified flow*** from ***web server*** to an ***application server*** *to* a ***database server*** to exiting the system)*:*

monitoring a processing time required for each application program to process the transaction received by each application server (as stated in col. 4, lines 19-21, col. 7, lines 58-67, ***objective*** of a ***load distribution policy*** is to ***minimize criterion*** such as the ***mean response time*** of a ***request***. FIG. 4 illustrates a ***sensor arrangement*** which are shared libraries or script ***components*** that ***intercept*** the ***actual processing time*** of a ***transaction request***. Each ***sensor*** is logically ***composed*** of two parts start part performs the ***correlation aspects*** of the ***monitoring***, and the end part forwards ***monitored data*** to collector 332 via ***measurement server*** 408)*;*

detecting a bottleneck relating to usage of at least one resource, wherein (as stated in col. 5, lines 21-25, each ***server*** is represented as a ***processor-sharing*** queue with one ***critical resource*** e.g., a ***CPU or disk***. The ***service demand*** of a ***request*** is ***obtained*** at a ***server***, which is the sum of ***processing times*** of the multiple passes of this ***request*** at the ***server***)*;*

each resource of the at least one resource is independently selected from the group consisting of a resource of at least one application server of the N application servers, a

resource related to input to the transaction, a resource of the database server, and a resource related to the transaction, wherein (as stated in col. 3, lines 5-17, col. 4, lines 44-46, For ***performance characterization***, the instrumentation ***data*** gathered in ***processing*** web ***transactions*** is ***classified*** according to ***user*** and ***transaction*** steps 204 and 206 and various transaction ***data*** is ***collected*** step 202. This transaction or instrumentation data includes ***data*** that ***identify*** transactions, ***classify*** transactions, identify ***requesters***, and ***quantify responsiveness*** of various ***resource components*** involved in ***processing*** the ***transaction***. ***Data collected*** at various components involved in processing the transactions is ***correlated*** by transactions and ***allocating resources*** efficiently to see that ***applications requirements*** are being ***satisfied and optimized.*** In addition to ***adjusting load balancing*** policies, the instrumentation ***data*** is also ***used*** in ***estimating*** and ***optimizing servers resources requirements*** step 210)*;*

said detecting is responsive to said monitoring having determined that the processing time for processing the transaction by I application servers of the N application servers is not within a predesignated permissible processing time range, and wherein I is at least 1 (as stated in col. 5, lines 23-27, lines 60-64, col. 4, lines 59-60, ***service demand*** of a ***request*** at a ***server*** is the sum of ***processing times*** of the ***multiple passes*** of this ***request*** at the ***server*** and expected ***response time*** may be described as the sum of response times at each of the three tiers, where average ***queuing time*** of the multi-tiered system then becomes the ***response time*** of the tiered system after adding to it some fixed "overhead" delays at ***non-bottleneck resources***

such as the ***fixed processing time*** at the ***load balancer. Average response time*** is then ***compared*** to the ***range of response time specified*** in the ***Service level Agreement*** SLA);

and removing the detected bottleneck (as stated in col. 5, lines 64-67, A mathematical ***optimization model*** is next ***formulated*** to find the ***optimal number*** of ***servers resources*** at each of the ***tiers***).

As to Claims 16, 25 and 33, *Garg teaches method, Load Control Server and Computer readable medium, of claims 15, 24 and 32, wherein M denotes a predesignated threshold number of application servers, wherein said detecting the bottleneck relating to usage of at least one resource comprises:*

identifying the at least one resource, and wherein said identifying the at least one resource comprises independently identifying each resource of the at least one resource as being (as stated in col. 5, lines 26-67, To ***obtain*** an ***estimation*** of the ***service demand*** at the ***application server*** tier E[S.sub.app], the ***relationship*** E[S.sub.app]=u.sub.appN.sub.app/.lamda..sub.app, where u.sub.app is the average ***utilization rate*** of the ***critical resource, .N.sub.app*** is the ***number of servers*** at the ***application server*** tier, .E [S.sub.app] is the average ***service demand);***

said resource of at least one application server of the N application servers if I is at least 1 and does not exceed M and if a processing time for processing another type of transaction by any application server of the N application servers is not within the predesignated permissible processing time range (as stated in col. 5, lines 65-67, col. 6,

lines 1-13, ***mathematical optimization*** model is ***formulated*** to ***find*** the ***optimal number*** of ***servers*** at each of the ***tiers***. The ***decision variables*** on which ***optimization*** are ***performed*** is the ***number of servers*** at ***each tier*** in the multi-tiered system. The objective function is the weighted sum of the number of servers at each tier, where the weights are the "costs" per server. The number of servers at each tier is constrained to be an integer greater than or equal to one. ***Optimization model*** has constraint E[R].ltoreq.SLA.sub.R, where SLA.sub.R is the ***response time limit*** such as 1 second ***required*** by the ***Service Level Agreement*** SLA)*;*

said resource related to input to the transaction if I is at least 1 and does not exceed M and if a processing time for processing another type of transaction by any application server of the N application servers is within the predesignated permissible processing time range (as stated in col. 6, lines 14-19, resulting ***mathematical optimization model*** has a ***linear objective*** function but a ***non-linear***, inequality-type constraint with integer-valued ***decision variables***. A ***concavity*** property of the ***average response time*** E[R] ***function*** is used with respect to the ***decision variables*** in formulating an ***efficient*** bounding ***procedure***)*;*

said resource of the database server if I exceeds M and if a processing time for processing another type of transaction by any application server of the N application servers is not within the predesignated permissible processing time range (as stated in col. 6, lines 19-24, bounding procedure ignores the integer-value requirements on the ***decision variables*** and ***solves*** the ***2-tiered problem***. The ***solution*** is then ***rounded***

to ***integer values***. Then the ***3-tiered problem*** is ***solved*** using the ***solution*** to the ***2-tiered problem*** and recursively to the general ***n-tiered problem***);

said resource related to the transaction if I exceeds M and if a processing time for processing another type of transaction by any application server of the N application servers is within the predesignated permissible processing time range (as stated in col. 6, lines 25-32, Once the ***server requirements*** have been ***estimated*** and ***optimized***, an ***assignment*** of ***applications to servers*** may be ***determined*** as a ***function*** of the ***optimal server requirements*** predicted in such a way ***communications delays*** are ***minimized*** and ***bandwidth capacity constraints*** are ***satisfied*** (step 212). The ***bandwidth capacity constraints*** are the ***actual bandwidth*** of the physical ***resources***).

As to Claims 17, 26 and 34, *Garg teaches method, Load Control Server and Computer readable medium, of claims 16, 25 and 33, wherein the method further comprises monitoring processing loads imposed on* (as stated in col. 6, lines 65-67, col. 7, lines 18-20, FIG. 3 is a ***functional*** block diagram of an example arrangement for ***gathering (monitoring) data*** to be used in ***analyzing resource requirements*** and ***allocations*** for ***applications*** hosted by a data center. The ***load balancer*** distributes ***transactions*** in a manner that ***minimizes response time*** and ***maximizes resource utilization)***:

resources of the N application servers, resources of the database server, and resources related to the transaction, and wherein on (as stated in col. 7, lines 36-

38, ***Collector*** 332 ***gathers instrumentation data*** pertaining to web ***transactions*** as the ***transactions are processed*** by each ***component from block*** 302 ***to database*** 316)*;*

said identifying each resource of the at least one resource comprises determining from the monitored processing loads that a high load specific to each resource of the at least one resource is imposed on each resource of the at least one resource (as stated in col. 7, lines 45-54, ***Analyzer-optimizer*** block 342 analyzes the ***correlated*** instrumentation ***data***, ***determines*** a ***desired configuration***, and initiates ***reconfiguration*** of the ***load balancer*** 306, ***servers*** and ***load balancers*** in the web server farm 308, and ***servers*** in ***application server farm*** 314 as may be ***desirable***. In an example embodiment, the analyzer-optimizer uses a ***queuing model*** to ***estimate*** and ***optimize server requirements*** of the ***applications*** based on mix of ***transaction types***, the ***volume of the different transaction types***, and a ***level of service*** that the data center is ***expected to provide)***.

As to Claims 18, and 27, *Garg teaches method, Load Control Server and Computer readable medium, of claims 17, and 26, wherein said determining that a high load is imposed on each resource of the at least one resource comprises* (as stated in col. 4, lines 8-10, Based on the ***user classifications***, ***workload mix***, and ***workload levels*** the ***load balancing policies*** may be ***adjusted)****:*

determining that a predesignated detection condition has occurred for each resource of the at least one resource a predesignated number of times, and wherein the

predesignated detection condition is that a predesignated usage parameter specific to each resource of the at least one resource is in a predesignated load range (as stated in col. 4, lines 34-43, The ***benefits*** of a ***sophisticated distribution policy*** based on ***user and URI classification*** may be sufficient to ***merit reassigning*** a ***session*** after the ***request*** in the ***session*** have been observed over some ***period of time***. This may provide a more ***accurate estimate*** of the ***sizes*** of subsequent ***requests*** in the session, if shortly into a ***session*** it is ***determined*** that the ***session*** is ***driven*** by a ***robot*** that will issue ***one type*** of ***request*** a ***large number of times,*** it might be worthwhile to ***reassign*** the ***session*** to a ***server dedicated*** to those ***types*** of ***requests***).

As to Claims 19, 28 and 35, *Garg teaches method, Load Control Server and Computer readable medium, of claims 17, 26, and 34, wherein said removing the detected bottleneck comprises eliminating the high load imposed on each resource of the at least one* (as stated in col. 7, lines 45-54, ***Analyzer-optimizer*** block 342 ***analyzes*** the ***correlated*** instrumentation ***data***, ***determines*** a ***desired configuration***, and initiates ***reconfiguration*** of the ***load balancer*** 306, ***servers*** and ***load balancers*** in the web server farm 308, and ***servers*** in ***application server farm*** 314 as may be ***desirable***. The ***analyzer-optimizer*** uses a ***queuing model*** to ***estimate*** and ***optimize server requirements*** of the ***applications*** based on ***mix*** of ***transaction types***, the ***volume (load) of the different transaction types***, and a ***level of service*** that the data center is ***expected to provide)***.

As to Claims 20, 29 and 36, *Garg teaches method, Load Control Server and Computer readable medium, of claims 19, 28, and 35, wherein said eliminating comprises executing in a predesignated sequence specific to each resource of the at least one resource as many of one or more predesignated load control processes as is necessary to eliminate the high load imposed on each resource of the at least one resource* (as stated in col. 4, lines 19-33, ***objective*** of a ***load distribution policy*** is to minimize some criterion such as the mean response time of a ***request*** by implementing several known load distribution policies. The ***load*** may be ***distributed*** based on a ***round-robin***, ***random***, ***least-work-remaining*** or ***size-based policy***. ***Sessions*** are considered in ***load balancing***. A ***session*** is a ***sequence*** of ***related*** Web ***requests***. In the example policy, the assignment for routing is performed once per session. To ***implement a minimizing-variance*** aspect of the ***size-based policy***, at the initial Web ***request*** of a ***session*** an ***estimation*** of the ***size*** of ***subsequent requests*** is made. ***Sessions*** comprising mostly ***small requests*** may be ***assigned*** to ***different servers*** from those ***comprising*** mostly ***large requests)***.

As to Claims 21, 30 and 37, *Garg teaches method, Load Control Server and Computer readable medium, of claims 20, 29, and 36, wherein a first resource of the at least one resource is a resource of a first application server of the N application servers, wherein said executing the predesignated sequence specific to the first resource comprises* (as stated in col. 12, lines 37-39, lines 65-67, ***mathematical*** formulation of the ***resource allocation problem*** (RAP) is described as follows. The ***application***

architecture requirements are represented by the following ***parameters***. The number of ***servers*** to be ***allocated*** to ***tier l*** is defined by ***N.sub.l***).

reducing an application program multiplicity of the first application server, and wherein said application program multiplicity on the first application server is defined as a maximum number of application programs to be executed concurrently on the first application server with respect to a plurality of transactions of the same type that were received by the first application server at the same time (as stated in col. 13, lines 1-39, col. 15, lines 62-67, The ***maximum*** and ***minimum attribute*** requirements are represented by two matrices VMAX and VMIN, where each element VMAX.sub.la and VMIN.sub.la represent the maximum and minimum ***level of attribute*** a for any ***server in tier l.*** The matrix T is defined to ***characterize the traffic*** pattern of the ***application***, where the ***element T.sub.li*** represents the ***maximum amount*** of ***traffic*** going from each ***server in tier l*** to each ***server in tier i.*** The numbers T.sub.01 and T.sub.10 represent the Internet traffic coming into and going out of each server in tier l. Using these traffic parameters, the ***total amount*** of incoming and outgoing ***traffic*** at each ***server*** in different tiers may be ***calculated***, denoted by *Tl.sub.l and TO.sub.l,* respectively. To reduce the ***number*** of ***binary variables x.sub.ls*** in the formulation, a ***feasibility matrix F*** is defined by Mixed Integer Programming problem, MIP2, to intelligently round the local optimal solution generated by the QP model. The MIP2 model ***defines*** the ***actual servers*** to ***allocate*** to the ***application***. The ***decision variables*** are the ***same*** as those in the ***original problem P0***).

As to Claims 22, 31 and 38, *Garg teaches method, Load Control Server and Computer readable medium, of claims 20, 29, and 36, wherein a first resource of the at least one resource is a resource of the database server, and wherein said executing the predesignated sequence specific to the first resource comprises reducing a priority level of a process for accessing the database* (as stated in col. 1, lines 58-60, col. 6, lines 49-62, for each application ***resource requirement*** may be determined as a function of the ***workload levels, service level*** metric associated with the application and ***subset of resources*** for each ***application***. The adjusting ***load balancing policies***, determining an ***allocation of resources***, and ***automatically reconfiguring*** may be repeated as often as deemed necessary to achieve ***desired levels*** of ***performance*** and ***efficiency***, ***reconfiguration*** tasks may include ***removing*** and installing application software, changing registry settings, editing of configuration files, and running a command to start the application software. The various ***scripts*** and ***sequences*** of operations needed for ***reconfiguration*** will vary according to the ***type of server*** and ***characteristics*** of the ***application software***).

As to Claim 23, *Garg teaches method, of claim 15, wherein an upper limiting processing time of the predesignated permissible processing time range is one standard deviation higher than an average processing time per transaction processed during peak processing loads during a predesignated period of time* (as stated in col. 4, lines 34-37, col. 5, lines 60-64, The benefits of a sophisticated ***distribution policy based*** on user and URI classification may be sufficient to merit ***reassigning*** a ***session*** after

the request in the session have been observed over some ***period of time*** and to approximate the ***average response time*** for a given number of servers at each tier, and an ***optimization process*** determines the minimum number of total ***servers*** required for the ***application average response time*** to be within ***time range*** of ***specific SLA***, ***service level agreement*** and ***average queuing time*** of the ***multi-tiered system*** then ***becomes*** the ***response time*** of the tiered system after ***adding*** to it some ***fixed "overhead" delays*** at non-bottleneck ***resources*** such as the ***fixed processing time*** at ***the load balancer***).

Conclusion

6. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

US Patent No. 6615253 to Bowman-Amuah; Michel K., US Patent No. 7069267 to Spencer, Jr., Herman, and US Patent No. 6950848 to Yousefi'zadeh, Homayoun, are cited for reference purpose only.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Muktesh G. Gupta whose telephone number is 571-270-5011. The examiner can normally be reached on Monday-Friday, 8:00 a.m. -5:00 p.m., EST.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, William C. Vaughn can be reached on 571-272-3922. The fax phone

number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

MG

/William C. Vaughn, Jr./

Supervisory Patent Examiner, Art Unit 2144